基于文本数据增强的生活满意度预测模型优化

陈佳婧[1,2] 胡丁鼎[1,2] 宋蕊[1,2] 谭诗奇[1,2] 李雨晴[1,2] 张胜楠[1,2] 朱廷劭[1,2*] 赵楠[1,2*]

[1]（中国科学院心理研究所 北京 100101）

[2]（中国科学院大学心理系 北京 100049）

摘要：

[目的]随着网络大数据以及机器学习的方法的发展，越来越多研究结合文本分析与机器学习来预测满意度。在建立生活满意度预测模型的研究中，针对获取大量有效的有标注数据困难的问题，本研究提出基于文本数据增强以优化生活满意度预测模型。

[方法]改编大连理工词典后，以 357 份生活现状描述为原始文本、生活满意度量表自评分为标注，经过 EDA 和回译进行文本数据增强，利用传统机器学习算法建立预测模型。

[结果]结果显示，大连理工词典改编后，各模型预测能力大大提高；数据增强后，仅在线性回归模型上观察到回译和 EDA 的提升作用。使用原始数据进行训练的岭回归模型预测值与实际值的皮尔逊相关系数最高，达 0.4131。

[结论]特征提取精度的提升可优化目前的生活满意度预测模型，但对于以词频为特征建立的生活满意度预测模型，基于回译和 EDA 进行的文本数据增强可能并不十分适用。

关键词：生活满意度；大连理工词典；文本数据增强；回译；EDA；机器学习

Optimization of a prediction model of life satisfaction based on text data augmentation

Chen Jiajing[1,2] Hu Dingding[1,2]Li Yuqing[1,2] Song Rui[1,2] Tan Shiqi[1,2] Zhang Shengnan[1,2] Zhu Tingshao[1] Zhao Nan[1*]

1 (Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101, China)
2 (University of Chinese Academy of Sciences, Beijing, 100049, China)

Abstract

[Objective] With the development of network big data and machine learning, more and more studies starting to combine text analysis and machine learning algorithms to predict individual satisfaction. In the studies focused on building life satisfaction prediction models, it is often difficult to obtain large amounts of valid and labeled data. This study aims at solving this problem using data augmentation and optimizing the prediction model of life satisfaction.

[Method] Using 357 life status descriptions annotated by self-rating life satisfaction scale scores as original text data. After preprocessing using DLUT-Emotionontology, EAD and back-translation method was applied and the prediction model was built using traditional machine learning

algorithms.
[Results] Results showed that (1) the prediction accuracy was largely enhanced after using the adapted version of DLUT-Emotionontology; (2) only linear regression model was enhanced after data augmentation; (3) rigid regression model showed the greatest prediction accuracy when trained by original data (r = 0.4131).
[Conclusion] The improvement of feature extraction accuracy can optimize the current life satisfaction prediction model, but the text data augmentation methods, such as back translation and EDA may not be applicable for the life satisfaction prediction model based on word frequency.
Keywords: Life Satisfaction; DLUT-Emotionontology; Text data augmentation;Back translation; EDA; Machine learning

1 引言

生活满意度（Life Satisfaction）指个体基于自身设定的标准对生活质量做出的主观评价，是个体对自己生活的综合判断(Papadopulos et al., 2007)。对生活满意度常用的测量方法为问卷调查法，如 Diener 的生活满意度问卷(Satisfaction with Life Scale,SWLS)。尽管这些量表具有较高信效度，但有研究者指出，使用问卷调查法测量生活满意度受到情境、记忆、被试参与意愿等因素的影响，生态效度不足。

近年来，随着机器学习的兴起，有研究者提出使用文本分析和机器学习建立满意度预测模型。其中被广泛使用的机器学习方法为有监督学习法，即对训练数据进行标注后再利用机器学习模型对结果变量进行预测，以达到较高预测精度(李静等, 2021；彭嘉丽等, 2021)。基于此，已有研究通过词典分词、情感分析等方法对个体主观幸福感(李昂等,2015;Wang et al., 2020)、环境满意度(Z. Wang et al., 2021)及电子产品满意度(Chatterjee et al., 2021)进行机器学习建模及预测，皮尔逊相关系数达到 0.3-0.5。然而，在目前的生活满意度研究中，获得有标注的数据困难度大，而较小的数据集可能导致模型出现过拟合问题，因此，如何获得大量有效的有标注数据是采用进行机器学习建立生活满意度模型亟需解决的问题之一。

当数据集较小时，数据增强技术可以使模型表现出更好的泛化能力和性能。数据增强是指通过对现有数据进行轻微修改产生副本或从现有数据创建新的合成数据来增加数据量的方法(Li, Hou, & Che, 2021)，被广泛应用于计算机视觉领域，如图像翻转和旋转，而后引入到自然语言处理（Natural Language Processing，NLP)，即文本数据增强。目前，数据增强在 NLP 中的应用较少，用

于文本数据增强的方法主要有以下几种：1.词汇替换：替换原始文本的某一部分，而不改变句子本身的含义(Ma & Langlang, 2020)。2. 简单数据增强（Easy Data Augmentation，EDA）：EDA 包含四个简单但功能强大的操作，即同义词替换、随机插入、随机交换和随机删除。对于给定的训练集，上述四种操作之一被随机选择并使用(Wei & Zou, 2019)。3.回译：使用机器翻译的方法来复述生成一段新的文本。一般步骤为先翻译成其他语言，再翻译回原始语言，从而在语义不变的情况下扩充数据量 (Lun, Zhu, Tang, & Yang, 2020; Ma & Langlang, 2020)。

文本数据增强被应用于传统模型和神经网络模型中，提升模型预测能力的效果良好。在传统模型中，Abdelrahman ElNaka（2021）等人采用回译等数据增强的方法扩大数据集，使随机森林、支持向量机等模型性能得到显著提升(ElNaka, Nael, Afifi, & Nada Sharaf, 2021)。而在神经网络模型中，Jun Ma（2020）等人也发现回译的数据增强方法可以提高深度学习分类模型对中文文本的分类能力(Ma & Langlang, 2020)；此外，Jiaqi Lun（2020）等人也证明在对简答题评分的深度学习模型中，文本数据增强有显著效果(Lun et al., 2020)。

综上，针对满意度机器学习建模中数据集过小的问题，本研究采用回译和 EDA 进行文本数据增强以扩大数据集，期望建立有较高预测能力的生活满意度预测模型。

2 方法

2.1 样本集

本研究共含 4 个样本集，分别为原始样本集、回译样本集、EDA 样本集 1、EDA 样本集 2。

原始样本集：在中国科学院大学随机选取研究生及博士生 392 人，要求被试填写生活满意度量表，算出总得分，并在 txt 文档中标注生活满意度量表的得分、性别、撰写一小段自评报告描述对自己目前状况的评价或感想。由 6 名筛选人员对 392 份文本数据进行筛选，筛选标准为自评报告字数大于等于 300 字、内容以第一人称叙述、不摘抄他人文本、有心情及生活经历等相关叙述。此外，6 名筛选人员还根据生活满意度量表的五个维度对现有文本进行评分（评分者一致性：Kendall’s W = 0.88，p<0.001)，若有 2 名以上评分者的评分与自评差距大于

5 分，剔除该文本。经筛选人员筛选后，剔除 35 份数据，最终合格数据共 357 份，其中男性被试 96 人，女性被试 261 人。

回译样本集：对训练集进行 6 次回译，将原始文本数据增加 6 倍，并与原始文本合并，得到 2499 个样本，成为回译样本集。

EDA 样本集 1：对每个原始文本，随机进行同义词替换、随机插入、随机交换或随机删除，改写比例（alpha）为 0.05 （Wei & Zou, 2019)。为与回译的训练集保持大小相同，将现有文本数据通过 EDA 改写增加 6 倍，得到 2499 个样本，与原始文本合并，成为 EDA 样本集 1。

EDA 样本集 2：根据 Wei & Zou（2019） 的研究，本研究原始样本的最佳改写倍数为 16，因此，以与 EDA 样本集相同的改写比例（0.05）将现有文本数据通过 EDA 增加 16 倍，与原始文本合并，得到 6069 个样本，成为 EDA 样本集 2。

2.2 工具

生活满意度问卷(Satisfaction with Life Scale, SWLS) 中文版（Diener, et al., 1985)：李克特七分量表，1 分代表非常不同意，7 分代表非常同意，共包含五个问题，将分数相加即作为被试的生活满意度总分。信度分析的结果显示，量表的 α 系数为 0.78, 折半信度为 0.70。表明生活满意度量表有较好的信度。

改编版大连理工词典：现有 21 个维度的大连理工情绪词典不能充分覆盖情绪表达词汇，我们首先在该词典的基础上增加含 5 个维度的微博常用情绪词库。进一步的，观察分词结果，发现大连理工词典和微博常用情绪词库中都缺少“否定词+情绪词”的复合词汇，而这样的表达出现频率较高。为提高特征提取精度，我们建立否定词库，从所有文本中找到所有不重复的“否定词+情绪词”的搭配。经 6 人小组讨论，在 1496 个复合情绪词中筛选出 1125 个无歧义的词语，进行情绪词库扩充，增加三类复合情绪词：P(Positive)，N(Negative)，Ne(Neutral)。最终形成 29 个维度的情绪词库。

2.3 数据处理

2.3.1 数据增强

回译：在 Python 中调用百度翻译 API 对训练集进行“中-英-中”，“中-法-中”，“中-德-中”，“中-俄-中”，“中-韩-中”，“中-日-中” 6 次回译。

EDA：改编自 Zhanlaoban（2019）的 github 程序在 Python 中，首先使用 jieba 分词包对原始文本进行分词，然后对每个文本的分词结果进行同义词替换、随机插入、随机交换或随机删除，每次 EDA 只采用一种改写方式。其中，同义词替换通过调用中文同义词包（synonyms），为选出的 n 个改写词（非停用词）分别找到一系列同义词，随机选择同义词进行词替换；随机插入通过为句子中的 n 个词（非停用词）找到随机的同义词，然后插入到句子的随机位置完成；随机交换通过随机选择句子中的两个词，进行位置交换，重复 n 次完成；随机删除通过以概率 p 删除文本中的词语完成。

2.3.2 特征提取

使用 jieba 分词包和改编版大连理工词典对清洗后的生活满意度文本数据进行分词，并删去分词后提供信息较少的词语。其后，基于改编版大连理工词典，计算每条文本 29 个情绪词维度的词频，得到 29 个特征。

2.3.3 模型建立及效果检验

为确保数据划分的一致性，每次划分训练集（80%）和测试集（20%）后，删去测试集中的增强文本，使只有训练集中的文本数据得到增强，得到增强训练集、未增强训练集和测试集。调用 python 的 scikit-learn 机器学习包，建立线性回归、岭回归、随机森林回归、决策树、支持向量回归和高斯过程回归六个模型。分别使用增强训练集和未增强训练集对各模型进行训练，得到 6 个增强模型和 6 个原始模型。将测试集特征值分别输入增强模型和原始模型，得到 12 个模型的测试集预测值，将该预测值与实际值进行皮尔逊相关分析，得到模型预测能力指标。以上过程重复 100 次。

3 结果

3.1 大连理工词典改编对模型预测能力的影响

使用原始样本集进行词典改编效果的检验。加入微博词库后，除岭回归外，各模型预测值与实际值的皮尔逊相关系数均有提高。从所有文本（含回译文本和改写文本）中筛选出所有不重复的“否定词+情绪词”的搭配并扩充词典后，各模型预测值与实际值的皮尔逊相关系数提高 0.09-0.13（表 1）。其中，最优预测模型为支持向量回归模型，在 100 次随机划分测试集与训练集的测试中，其预测

值与实际值的皮尔逊相关系数最高为 0.5971。

表 1. 词典改编对各模型预测能力的影响（r）

词典	测试模型					
	线性回归	岭回归	随机森林回归	决策树	支持向量回归	高斯过程回归
原版大连理工词典	0.22590	0.31605	0.26790	0.09389	0.28010	0.27745
微博词库+大连理工词典	0.23618	0.29207	0.30447	0.12607	0.29755	0.31050
改编版大连理工词典	**0.34121**	**0.42394**	**0.41195**	**0.20791**	**0.41216**	**0.40650**

注：表中数值为皮尔逊相关系数（r），由 100 次随机划分测试集与训练集的模型测试结果求平均值所得。

3.2 数据增强对模型预测能力的影响

由于采用改编版大连理工词典进行特征提取时，各模型的预测能力最佳，因此后续特征提取都采用改编版大连理工词典进行。

将各模型预测值与生活满意度自评进行皮尔逊相关分析，结果显示，在线性回归、岭回归、随机森林回归、决策树、支持向量回归和高斯过程回归六个模型中，数据增强的效果不同（表 2）。

表 2. 数据增强对各模型在测试集中预测能力的影响（r）

训练集	测试模型					
	线性回归	岭回归	随机森林回归	决策树	支持向量回归	高斯过程回归
回译样本集						
未增强训练集	0.33956	0.40119	0.40491	0.21851	0.39848	0.39355
增强训练集	**0.37915**	0.39426	0.37550	0.20681	0.39628	0.38705
EDA 样本集 1						
未增强训练集	0.34602	0.41132	0.40366	0.21850	0.41204	0.40622
增强训练集	0.32290	0.37873	0.35401	0.18785	0.39502	0.37915
EDA 样本集 2						
未增强训练集	0.34548	0.41046	0.41191	0.21952	**0.41310**	0.40741
增强训练集	**0.35562**	0.37163	0.37523	0.21334	0.37683	0.37782

注：表中数值为皮尔逊相关系数（r），由 100 次随机划分测试集与训练集的模型测试结果求平均值所得。

如表 1 和图 1 所示，采用回译进行数据增强时，只有线性回归模型的皮尔逊相关系数提高 0.04，其它各模型的皮尔逊相关系数均降低（0.002-0.029）。采用 EDA 进行 6 倍数据增强时，各模型的皮尔逊相关系数都降低（0.017-0.049）。采用 EDA 进行 16 倍数据增强时，只有线性回归模型的皮尔逊相关系数提高 0.010，其他各模型的皮尔逊相关系数都降低（0.006-0.038）。

在所有训练集的 100 次训练结果中，岭回归模型在使用未增强样本集进行训练后，皮尔逊相关系数最高（r=0.41647）。

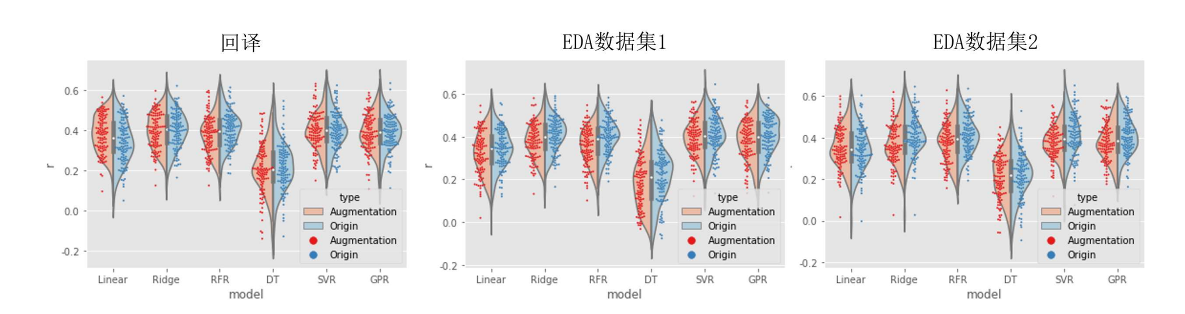


图 1. 数据增强前后各模型在测试集中预测能力的变化

（注：Linear=线性回归，Ridge=岭回归，RFR=随机森林回归，DT=决策树，SVR=支持向量回归，GPR=高斯过程回归）

4 讨论

随着信息技术的不断发展，如今人们会在各种各样的平台上分享自己的想法，这些均以文本的形式展现，通过分析这些文本，有望对人们的生活满意度进行预测。在本研究中，针对满意度机器学习建模中数据集过小的问题，本研究采用回译和 EDA 进行文本数据增强以扩大数据集，期望建立有较高预测能力的生活满意度预测模型。结果表明，改编版大连理工词典提高了传统机器学习模型的预测能力；数据增强中，回译和 EDA 的增强方法对线性回归模型的预测能力有提升作用。

为大连理工词典增添微博词库及复合情绪词后，在未经数据增强时，支持向量回归对个体生活满意度进行预测与个体自评生活满意度分数之间的最优相关系数能够达到 0.5。以往研究表明，社会与人格心理学领域中，不同测量工具之间的相关系数位于 0.39-0.68 之间(李昂 等人, 2015)。此外，也有研究使用机器学习算法预测个体主观幸福感，其最优取值结果介于 0.27~0.60(李昂

等人，2015)。在本研究中，模型预测生活满意度的相关系数能够达到 0.5 左右，表明模型的效果良好。这个结果提示我们，结合文本分析以及机器学习算法对个体生活满意度进行预测的方法较为可靠。

本研究使用回译和 EDA 的方法对文本数据进行增强，发现增强后，各模型的表现不同，仅有线性回归模型在数据增强后表现出预测能力的提升。这些表现与以往文献中的不尽相同。在以往文献中，有研究者使用回译、EDA、预训练语义模型等数据增强的方式将数据扩充后使用不同的机器学习模型来预测，发现线性回归以及支持向量机模型在数据增强后变好，而随机森林模型在数据增强后变差(Ansari, Garg, & Saxena, 2021)。但是也有研究提取词向量为特征，分别使用支持向量机、随机森林以及神经网络对未增强以及回译增强后的文本数据集进行学习，发现相较于未增强的数据，三种机器学习算法使用回译增强后的数据的学习效果均变得更好(ElNaka et al., 2021)。对于此，有研究者认为，数据增强的效果会随着数据增强方式的不同而发生改变——弱增强往往能够提高预测精度而强增强可能会减弱预测精度(Min et al., 2021)。此外，Raghunathan 等人(2020)发现，使用增强后的数据进行模型训练会产生更小的稳健误差(robust error)，但是可能会产生更大的标准误差(standard error)。此外，在自然语言处理相关研究中，运用数据增强方法来进行数据预处理的研究大多都是基于深度学习模型来进行预测，这可能是由于深度学习模型非常依赖于大量的数据量来避免过拟合的问题的原因(Wen et al., 2020; Shorten, Khoshgoftaar, & Furht, 2021)。这些均说明数据增强为模型带来的效果可能依赖于增强方式、特征提取与模型特征。因此，在今后的研究中，可以尝试使用情感分析、词向量计算等方式进行特征提取，或使用深度学习模型来进行训练以及预测。

5 结论

改编大连理工词典后，各模型预测能力均有大幅提高，说明特征提取精度的提升可以提高目前生活满意度预测模型的预测能力。但文本数据增强后，仅在线性回归模型上观察到数据增强对模型预测能力的提升，说明对于以词频为特征建立的生活满意度预测模型，基于回译和 EDA 进行的文本数据增强可能并不十分适用。

参考文献

李昂,郝碧波,白朔天 & 朱廷劭.(2015).基于网络数据分析的心理计算:针对心理健康状态与主观幸福感. *科学通报*(11),994-1001.

李静, 刘德喜, 万常选, 刘喜平, 邱祥庆, 鲍力平, & 朱廷劭. (2021). 社会网络用户心理健康自动评估研究综述. *中文信息学报*, *35*(2), 19 - 32.

彭嘉丽, 赵英亮, & 王黎明. (2021). 基于深度学习的视频异常行为检测研究. *激光与光电子学进展*, 58(6), 51 - 61.

Ansari, G., Garg, M., & Saxena, C. (2021). Data Augmentation for Mental Health Classification on Social Media. *arXiv preprint* arXiv:2112.10064.

Chatterjee, S., Goyal, D., Prakash, A., & Sharma, J. (2021). Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application. *Journal of Business Research*, *131*(January 2020), 815–825.https://doi.org/10.1016/j.jbusres.2020.10.043

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of personality assessment*, *49*(1), 71-75.

ElNaka, A., Nael, O., Afifi, H., & Sharaf, N. (2021). AraScore: Investigating Response-Based Arabic Short Answer Scoring. *Procedia Computer Science*, 189, 282-291.

Li, B., Hou, Y., & Che, W. (2021). Data Augmentation Approaches in Natural Language Processing A Survey. *Journal of LATEX Templates*.

Lun, J., Zhu, J., Tang, Y., & Yang, M. (2020). Multiple Data Augmentation Strategies for Improving Performance on Automatic Short Answer Scoring. *The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence*, 13446-13453.

Ma, J., & Langlang. (2020). Data Augmentation For Chinese Text Classification Using BackTranslation. *Journal of Physics: Conference Series, 1651 (2020) 012039*.

Min, Y., Chen, L., & Karbasi, A. (2021, December). The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. *In Uncertainty in Artificial Intelligence* (pp. 129-139).

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J., & Liang, P. (2020). Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint* arXiv:2002.10716.

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep

learning. *Journal of big Data*, 8(1), 1-34.

Wang, Y., Wu, P., Liu, X., Li, S., Zhu, T., & Zhao, N. (2020). Subjective well-being of Chinese Sina Weibo users in residential lockdown during the COVID-19 pandemic: Machine learning analysis. *Journal of Medical Internet Research*, *22*(12), 1–13. https://doi.org/10.2196/24775

Wang, Z., Zhu, Z., Xu, M., & Qureshi, S. (2021). Fine-grained assessment of greenspace satisfaction at regional scale using content analysis of social media and machine learning. *Science of the Total Environment*, *776*, 145908.

Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *arXiv*.

Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., & Xu, H. (2020). Time series data augmentation for deep learning: A survey. *arXiv preprint* arXiv:2002.12478.

ZaneMuir. (2017). DLUT-Emotionontology. https://github.com/ZaneMuir/DLUT-Emotionontology

Zhanlaoban.(2019).EDA_NLP_for_Chinese. https://github.com/zhanlaoban/EDA_NLP_for_Chinese/commits/master.

作者贡献声明

陈佳婧：研究设计，数据清洗，代码实现，数据分析，方法、结果及结论部分撰写，论文整合

胡丁鼎：研究设计，数据清洗，代码实现，数据分析，讨论部分撰写

宋　蕊：研究设计，数据清洗，代码实现，数据分析，引言部分撰写

谭诗奇：研究方案讨论，数据清洗，引言部分撰写

李雨晴：研究方案讨论，数据清洗，数据分析，方法部分撰写，格式修订

张胜楠：研究方案讨论，数据清洗，引言部分撰写

朱廷劭、赵楠：原始数据收集，技术分析支持，论文指导

致谢

诚挚感谢中山大学计算机学院国家超级计算广州中心卢圣有同学的技术支持。